Privacy Over RDF datasets

Sara Taki, Cédric Eichler, Benjamin Nguyen

Privacy Over RDF datasets

INSA Centre Val de Loire, Laboratoire d'Informatique Fondamentale d'Orléans, France
Sara Taki, Cedric Eichler, Benjamin Nguyen

Context and Motivation

The Resource Description Framework (RDF) is a standard model at the core of linked data. An RDF data set is a set of triples (subject-predicate-object) which form a labeled directed graph with an underlying semantic. The use of Linked Data is increasing, and thus privacy in such data sources is becoming an issue [1]. Indeed, publishing or querying graph data may result in disclosure of sensitive information and therefore to the violation of individual privacy.

Objectives

- Study privacy over RDF datasets
- Adapt differential privacy (DP) to edge-labeled directed graphs with an underlying semantic

Differential Privacy

A randomized mechanism K: $D^n \rightarrow \mathbb{R}^d$ preserves (ϵ)-differential privacy [2] if for any pair of neighboring databases (x, y) $\in (D^n)^2$ and for all sets S of possible outputs:
Pr[K(x) $\in$ S] $\leq e^{\epsilon}$ Pr[K(y) $\in$ S] .

Achieving Differential Privacy can be done by adding an appropriate amount of randomized noise to the query results.

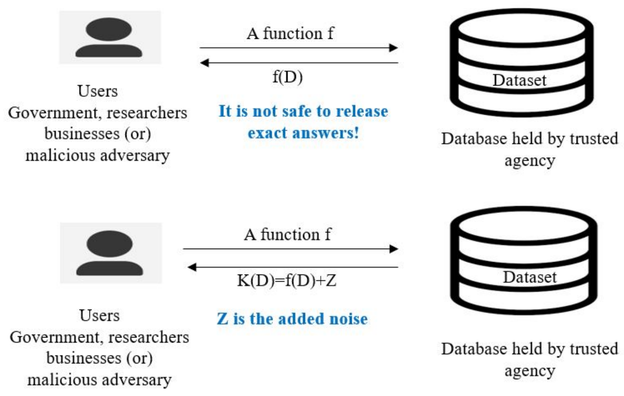


Figure 1: Achieving Differential Privacy

Calibration: for f : $D^n \rightarrow \mathbb{R}^d$ the global sensitivity of f with regard to a distance m over D^n is

$$\Delta_m f = \max_{(D,D') \in D^n s.t. m(D,D')=1} \| f(D) - f(D') \|_1$$

where $\|\|_1$ denotes the L1 norm

Possible noise, the Laplacian mechanism: In order to publish f(D) where f :$D^n \rightarrow \mathbb{R}$ while satisfying ϵ-DP, one may publish

$$K(D) = f(D) + Lap(\Delta f/\epsilon)$$

where Lap (Δf /ϵ) represents a random draw from the the Laplace distribution centered at 0 with scale Δf /ϵ

Example f: count the number of persons with flu
Δf=1, f(D)= 3, Laplace Mechanism outputs 3 + Lap(1/ϵ)

Name	Flu
Alice	0
Bob	1
Joe	1
Umeko	0
Carine	1

Figure 2: COUNT Query

Adapt differential privacy (DP) to edge-labeled directed graphs

1. Define distances and neighborhood (adjacency) between graphs
   - Node-DP (d_n): two graphs are neighbours if they differ by at most one node and all of its incident edges
   - Outedge-DP (d_o): two graphs are neighbours if they differ by the outlinks (outedges) of a chosen node
   - QL-out-edge-DP (d_{ql}) [3]: two graphs are neighbours if they differ by the sensitive outedges of one node. It relies on the definition of the set QL of edge types (labels) that are considered sensitive.

2. Compute query sensitivity Depending on the notion of neighborhood. For example, with f outputing the maximum out-degree in the graph
   - Node-DP: $\Delta_{d_n} f = n$
   - Outedge-DP: $\Delta_{d_o} f = n - 1$
   - QL-out-edge-DP: $\Delta_{d_{ql}} f = n - 1$

Problem:
- GS is high and sometimes unbounded (when considering family of all graphs)
- Adding noise which is proportional to n means adding too much noise
- Destroy the utility regardless of the used utility metric

Approach

Graph Projection: transform the original graph G into a graph of bounded degree, out-degree, or QL-out-degree.
Let g = family of all graphs
g^D= family of graphs with maximum degree D
g_o^D = family of graphs with maximum out-degree D
g_{QL}^D= family of graphs with maximum QL-out-degree D for a given QL $\subseteq L$

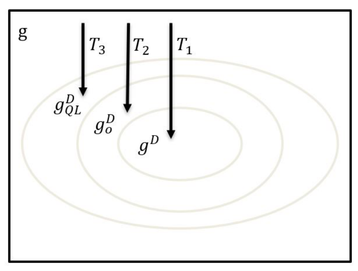
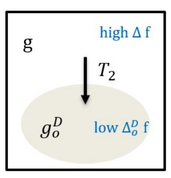


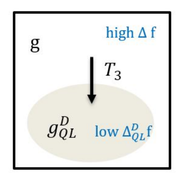
Figure 3: Proposed Projection Framework

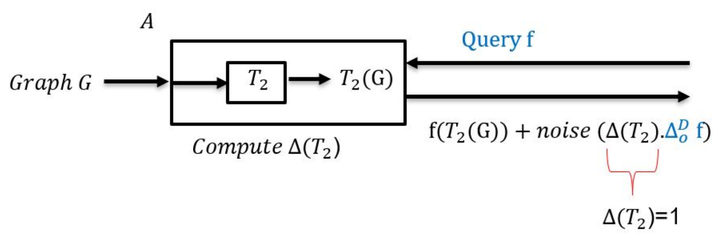
From Privacy on Bounded Out-Degree (resp. Bounded Ql-Out-Degree) Graphs to Privacy on arbitrary Graphs

Δf=global sensitivity of f over g
$\Delta_o^D f$=global sensitivity of f over g_o^D
$\Delta_{QL}^D f$=global sensitivity of f over g_{QL}^D

Outedge Privacy with T_2
Input: Algorithm B that is Outedge-DP over g_o^D
Output: Algorithm A that is Outedge-DP over g
QL-Outedge Privacy with T_3
Input: Algorithm D that is Ql-Outedge-DP over g_{QL}^D
Output: Algorithm C that is Ql-Outedge-DP over g

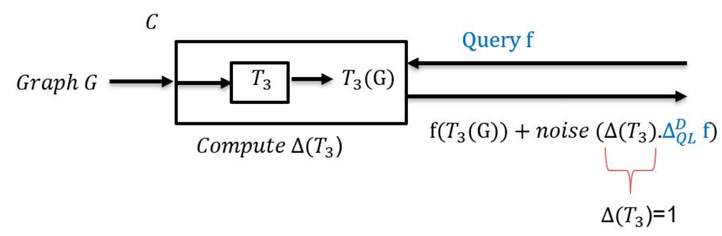






Theorem:[4]

$$\Delta(f \circ T) \leq \Delta_o^D f . \Delta T_2$$



$$\Delta(f \circ T) \leq \Delta_{QL}^D f . \Delta T_3$$

Conclusion And Future Work

- New approach based on graph projection to adapt differential privacy to edge-labeled directed graphs –e.g. RDF graphs– while reducing the amplitude of the randomized noise.
- Future work is mainly dedicated to experimentation on large datasets.

Acknowledgements

Work supported by the French National Research Agency, under grant ANR-18-CE23-0010

[1] R. Delanaux, A. Bonifati, M-C. Rousset, and R. Thion. *Query-Based Linked Data Anonymization.* Proceedings of the 17th International Semantic Web Conference, 2018.

[2] C. Dwork. *Differential Privacy.*Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 4052)

[3] J. Reuben.*Towards a differential privacy theory for edge-labeled directed graphs.* Proceedings of SICHERHEIT, 2018.

[4] SP. Kasiviswanathan, K. Nissim, S. Raskhodnikova, A. Smith. *Analyzing graphs with node differential privacy.* Proceedings of the Theory of Cryptography Conference, 2013.